Joint genotyping on the fly: Identifying variation among a sequenced panel of inbred lines

Eric A. Stone[1]

Department of Genetics and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27603, USA

High-throughput sequencing is enabling remarkably deep surveys of genomic variation. It is now possible to completely sequence multiple individuals from a single species, yet the identification of variation among them remains an evolving computational challenge. This challenge is compounded for experimental organisms when strains are studied instead of individuals. In response, we present the Joint Genotyper for Inbred Lines (JGIL) as a method for obtaining genotypes and identifying variation among a large panel of inbred strains or lines. JGIL inputs the sequence reads from each line after their alignment to a common reference. Its probabilistic model includes site-specific parameters common to all lines that describe the frequency of nucleotides segregating in the population from which the inbred panel was derived. The distribution of line genotypes is conditional on these parameters and reflects the experimental design. Site-specific error probabilities, also common to all lines, parameterize the distribution of reads conditional on line genotype and realized coverage. Both sets of parameters are estimated per site from the aggregate read data, and posterior probabilities are calculated to decode the genotype of each line. We present an application of JGIL to 162 inbred *Drosophila melanogaster* lines from the *Drosophila* Genetic Reference Panel. We explore by simulation the effect of varying coverage, sequencing error, mapping error, and the number of lines. In doing so, we illustrate how JGIL is robust to moderate levels of error. Supported by these analyses, we advocate the importance of modeling the data and the experimental design when possible.

[Supplemental material is available for this article.]

With recent advances in high-throughput sequencing technology, sequencing a genome can be accomplished affordably and with great speed. Consequently, it has become possible to survey genomic variation at unprecedented resolution, and many such studies are under way (e.g., The 1000 Genomes Project Consortium 2010; Li et al. 2010; Yi et al. 2010). Despite the ease with which massive quantities of sequence reads can now be obtained, interpretation of the data is nontrivial. Errors in sequencing and assembly may masquerade as genomic variation, and biases in these processes can complicate the resolution of heterozygous genotypes. By necessity, novel analysis tools are being developed with these challenges in mind.

In particular, for population genomic studies in which multiple individuals are sequenced, joint analysis has proven to be a successful strategy for resolving genotypes and identifying segregating genetic variation (The 1000 Genomes Project Consortium 2010; Le and Durbin 2010; DePristo et al. 2011; Li 2011). Provided that the sequenced individuals are related in some way (e.g., a family or a population sample), joint inference can be used to leverage the dependence between individual genotypes. The rationale behind joint inference is intuitive: Knowing that one individual has an A allele increases the likelihood that a relative, however distant, also has an A. This feature is particularly useful when many individuals are sequenced at low coverage, because the aggregate coverage still permits an accurate description of variation at the population level. Knowledge of the population, in turn, may improve the accuracy with which each individual's genotype can be resolved.

Whereas population genomics considers individuals within a population, cancer genomics targets a population of cells within an individual. For example, the latter might apply high-throughput sequencing technology to a tumor sample toward identifying variation with respect to the normal genotype of an affected individual. Although the technologies facilitating population and cancer genomics are largely the same, each field relies on distinct tools for data analysis. In particular, methods specific to cancer genomics must respect the fact that allele frequencies within a somatic tumor sample need not follow the germline expectation of strict homozygosity (0%/100%) or heterozygosity (50%). Several approaches doing so are in use already (Koboldt et al. 2009; Goya et al. 2010), with several more reported to be under development (Ding et al. 2010).

Each of the methods referenced above shares a common motivation in applications to human genomics; however, high-throughput sequencing technology is transforming the genomic studies of other species as well. Sometimes the methods developed in response to human studies apply equally well to other organisms, but this is not always the case. For example, an often distinguishing characteristic of non-human genomics is that a strain is studied rather than a single individual. By design, the individuals comprising a strain are genetically very similar, giving meaning to the concept of a strain's genotype. On the other hand, in most cases, a strain is not isogenic, implying that some degree of genetic variation remains. In that sense, a strain is not unlike a tumor sample, because while the strain is viewed as an individual, there is a population that underlies it.

It is becoming increasingly common for multiple strains to be studied simultaneously, for example, to map the quantitative trait loci responsible for a particular phenotype (Aylor et al. 2011; Cao et al. 2011). A prerequisite for such studies is to identify and characterize genetic variation, which is consistent with the goals of population genomics. High-throughput sequencing promises the unbiased investigation of large strain panels, but only once each strain's genotype has been accurately resolved. Following the logic of human population studies, if the strains comprising a panel are

[1] **Corresponding author.**
E-mail eric_stone@ncsu.edu.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.129122.111.

22:966–974 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

related in some way, there should be some benefit in simultaneously decoding the genotypes of each strain. But, following the lesson of cancer genomics, it must be appreciated that a strain is a population and not an individual.

This study considers the problem of joint genotypic inference on a large panel of strains. The work was motivated by challenges encountered in the *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al. 2012), a collection of 192 inbred *D. melanogaster* lines ("strains"). The DGRP lines were derived from a common Raleigh, NC population by 20 generations of full-sib mating (Fig. 1). The offspring of the sib pair at generation 20 and their descendants in perpetuity comprise the inbred line. Sequencing of the DGRP lines was done primarily on the Illumina GAII platform, and crucially what was sequenced for each line is DNA from a large pool of flies (between 500 and 1000 flies) (see Fig. 2). From these DNA pools, we sought to obtain genotypes simultaneously for each inbred line while respecting the fact that each line is itself a population. Because we know precisely how the lines are related, we were able to incorporate the DGRP experimental design. In what follows we describe our probabilistic model and its implementation as the Joint Genotyper for Inbred Lines (JGIL). We demonstrate how JGIL performs on data from the DGRP, and we report the results of a simulation study designed to interrogate how performance varies with coverage, sequencing error rate, mapping error rate, and the number of lines in the analysis.

Genotypes for inbred lines

The effect of inbreeding is to reduce the genetic variation within each DGRP line so that the majority of sites are fixed. At such positions, the genotype of the line is simply the genotype of any one individual; however, for sites that harbor residual variation, what is meant by the line's genotype is unclear. Our model of residual variation and our corresponding definition of "genotype" are directly based on the inbreeding scheme (Figs. 2, 3; Table 1).

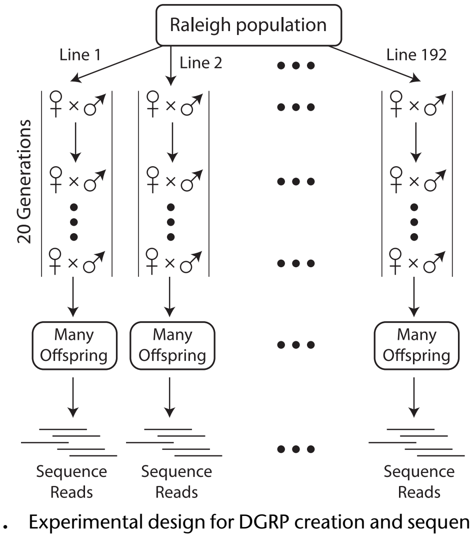

Figure 1. Experimental design for DGRP creation and sequencing. Each DGRP line was founded by a mated female collected from the Raleigh, North Carolina Farmer's Market. Each subsequent generation was created by crossing a pair of male and female progeny from the previous generation. The DGRP lines were produced by 20 generations of full-sib inbreeding. For each line, high-throughput sequencing was performed on DNA that was extracted from a pool of 500–1000 flies.

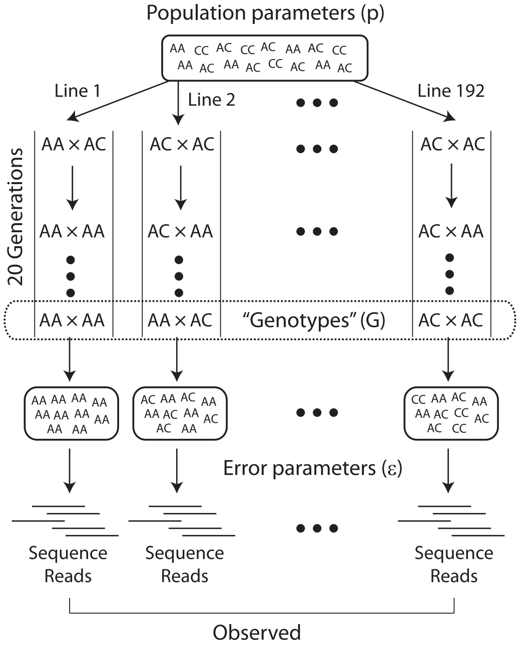

Figure 2. Schematic of model for site-specific joint genotypic inference. The philosophy of JGIL is for the model to mirror the experimental design. The founding females and their implicit mates are sampled from a common population that is described by population-level parameters. The experimental design specifies the probabilistic dependency between the genotypes of these initial flies and those of the pooled inbred samples that are ultimately sequenced. Technical and bioinformatic errors are modeled as a site-specific phenomenon that applies equally to all lines. This allows "line effects" (i.e., genotypes) and "nucleotide effects" (i.e., errors) to be disentangled.

Specifically, we take advantage of the fact that residual variation in the line is reflective of residual variation in the sib pair mated in generation 20 (excluding the contribution of new mutations) by relating the genotype of the line to those of this sib pair. We make two simplifying assumptions, namely, that (1) at most, two alleles are present among the progenitors of any one line (i.e., generation 0, although we permit more than two alleles to be segregating in the initial population); and (2) the transmission of alleles from the G20 sib pair to their inbred line (see Fig. 2) is not distorted. These assumptions allow us to define 22 discrete "line genotypes" (see Table 1) and assign probabilities to each (see Methods) based on the inbreeding design and the composition of the Raleigh population (as given by $\mathbf{p}$ below).

Joint genotypic inference

Restricting attention to one genomic site, consider m lines with genotypes $G_1, \ldots, G_m$ (collectively $\mathbf{G}$) and respective sets of covering reads $R_1, \ldots, R_m$ (collectively $\mathbf{R}$). Let $\boldsymbol{\theta}$ be a parameter vector that models aspects of the sequencing process and of the individuals themselves. The advantage of joint inference lies in using all of the data $R_1, \ldots, R_m$ to obtain a better estimate of $\boldsymbol{\theta}$ than would be possible when considering each individual separately.

We analyze each genomic position separately, and the values taken by the parameter vector $\boldsymbol{\theta}$ are unique to each position. The vector $\boldsymbol{\theta} = (p_A, p_C, p_G, p_T, \varepsilon_A, \varepsilon_C, \varepsilon_G, \varepsilon_T) = (\mathbf{p}, \boldsymbol{\varepsilon})$ includes eight parameters. The entries of $\mathbf{p}$ sum to 1 and describe the position-specific nucleotide frequencies in the common population from which the

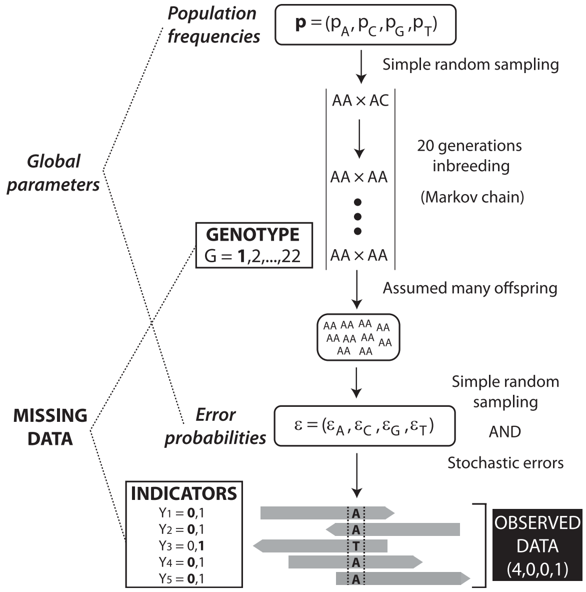

Figure 3. Detailed model and estimation framework for one line at one site. Frequencies of A, C, G, and T in the population govern the distribution of parental genotypes in generation 0 (AA × AC in the figure). The distribution of parental genotypes at generation 20, conditional on the genotypes at generation 0, is specified by the Markov chain described in Methods. It is assumed that this cross produces many offspring in the absence of segregation distortion so that the nucleotide frequencies among the offspring match those of the parents. The sequencing reads, which are the observed data, are composed of a random sample of these nucleotides (with replacement) along with errors whose frequencies are given by ε. The unobserved, or missing, data are composed both of the parental genotypes at generation 20 (here $G = 1$ for AA × AA) (cf. Table 1) and indicators ($Y_i = 0, 1$) that record the error status of each read. For example, $Y_3 = 1$ in the figure because the T is an error, which is clear when the "genotype" $G = 1$ is known, provided mutation is precluded. While each line has its own observed and missing data, the global parameters $\mathbf{p}$ and ε are common to all lines.

inbred lines were derived. The entries of $\boldsymbol{\varepsilon}$ describe the position-specific probabilities of obtaining a read with an erroneous base of A, C, G, or T, respectively. Thus, $\boldsymbol{\theta}$ parameterizes two generative probability distributions for each line i: (1) $\Pr(G_i|\boldsymbol{\theta}) = \Pr(G_i|\mathbf{p})$, which gives the distribution of the line's genotype G_i given the population frequencies $\mathbf{p}$; and (2) $\Pr(R_i|G_i, \boldsymbol{\theta}) = \Pr(R_i|G_i, \boldsymbol{\varepsilon})$, which gives the distribution of the line's reads R_i given its genotype G_i and error profile $\boldsymbol{\varepsilon}$. Note that while ε is specific to each genomic site, it does not vary with the position that reads cover a given site. Thus, $\boldsymbol{\varepsilon}$ models mapping error and context-dependent sequence error, but not variability in sequence quality due to read position.

Estimation strategy

We estimate $\boldsymbol{\theta}$ by maximum likelihood. The joint probability of the read data $\mathbf{R}$ across all lines is given by

$$\Pr(\mathbf{R} = \mathbf{r}|\boldsymbol{\theta}) = \prod_{i=1}^{m} \sum_{g=1}^{22} \Pr(R_i = r_i|G_i = g, \boldsymbol{\varepsilon})\Pr(G_i = g|\mathbf{p})$$

and we seek $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{R}) = \text{argmax}_{\boldsymbol{\theta}} \Pr(\mathbf{R}|\boldsymbol{\theta})$. We do so using the Expectation-Maximization algorithm (Dempster et al. 1977). We envision two layers of missing data, namely, the genotypes $\mathbf{G}$ and error indicator variables for each read, which we will call $\mathbf{Y}$ (see Fig. 3; Methods). While the algorithm only guarantees the discovery of a local maximum of the likelihood surface, for this application it appears as though the global optimum is being found whenever the initial choice of $\boldsymbol{\theta}^0$ is reasonable. Nearly exact analytical solutions to the maximization step (i.e., $\boldsymbol{\theta}^1 = \text{argmax}_{\boldsymbol{\theta}} \text{E}_{\mathbf{G},\mathbf{Y}|\mathbf{R},\boldsymbol{\theta}=\boldsymbol{\theta}^0}[\log \Pr(\mathbf{R}, \mathbf{G}, \mathbf{Y}|\boldsymbol{\theta})]$) are given in the Methods section. An analysis of the approximation error is given in Supplemental Figure 1.

Maximum a posteriori genotypes

Upon obtaining a maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$, the posterior probability of each genotype can be computed and is proportional to $\Pr\left(G_i = g|R_i = r_i, \hat{\boldsymbol{\theta}}\right)$. Maximum a posteriori (MAP) genotypes are assigned as $\hat{g}_i = \text{argmax}_g \Pr\left(G_i = g|R_i = r_i, \hat{\boldsymbol{\theta}}\right)$.

Genotype quality scores

Quality scores are assigned to each called genotype according to the *phred* scale (Ewing and Green 1998). Specifically, if the probability of the MAP genotype is P, then we report its quality as $Q = -10\log_{10}(1 - P)$.

Results

We begin with a summary of how JGIL performs on the Illumina data generated as part of the DGRP project. We evaluated this performance by means of three independent comparisons. First, because 29 of the DGRP lines considered here were also sequenced on the 454 Life Sciences (Roche) platform, we were able to compare the consistency of genotype calls across technologies. Second, because a pair of duplicate lines was sequenced, we were able to compare the consistency of genotype calls across biological/tech-

Table 1. Encoding of the 22 possible "genotypes," summarized as 10 states

	1	2	3	4	5			6			7			8			9			10		
	A	C	G	T	A+C (M)			A+G (R)			A+T (W)			C+G (S)			C+T (Y)			G+T (K)		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	4	0	0	0	3	2	1	3	2	1	3	2	1	0	0	0	0	0	0	0	0	0
C	0	4	0	0	1	2	3	0	0	0	0	0	0	3	2	1	3	2	1	0	0	0
G	0	0	4	0	0	0	0	1	2	3	0	0	0	1	2	3	0	0	0	3	2	1
T	0	0	0	4	0	0	0	0	0	0	1	2	3	0	0	0	1	2	3	1	2	3

Each entry M_{ij} in the 4 × 22 matrix **M** records the number of i alleles (A = 1, C = 2, G = 3, T = 4) in genotype j. The 22 columns of **M** are what we call "genotypes" and are further collapsed into 10 states that describe the nucleotides segregating within a line at a site. The states are (1) A, adenine; (2) C, cytosine; (3) G, guanine; (4) T, thymine; (5) M, amino [A and C]; (6) R, purine [A and G]; (7) W, weak [A and T]; (8) S, strong [C and G]; (9) Y, pyrimidine [C and T]; (10) K, keto [G and T].

nical replicates. Third, because a previous study used Sanger sequencing to target five ~1-kb regions of the *X* chromosome in all of the DGRP lines, we were able to compare a subset of the JGIL genotype calls to this gold standard.

Analysis of the DGRP data

JGIL was used to genotype the 162 inbred *D. melanogaster* lines in Freeze 1 of the DGRP for which Illumina sequences were available (Mackay et al. 2012). A second set of 40 lines, including 29 of these 162, was sequenced on the 454 platform, and from these JGIL produced a second set of genotype calls. The 29 lines sequenced on both technologies form the basis of our first comparison.

Specifically, for each of the 29 lines, we genotyped ~119 Mb of euchromatic DNA on the five major chromosome arms (*X*, *2L*, *2R*, *3L*, *3R*). For each line at each site, the MAP genotype was reported as one of the 10 states in Table 1 provided it received a quality score of 20 and there was greater than zero coverage; otherwise, the genotype was assigned an "N." Genome-wide, when a genotype was reported for both technologies, the agreement between them was 99.97%. In only 0.33% of cases, the Illumina data produced an N when the 454 data yielded a called genotype. Owing in part to lesser coverage, the 454-based call was an N in 2.09% of cases when a call was made from Illumina data.

The SNP sets from both technologies were also highly concordant. Both identified roughly the same number of SNPs (~2.8 million), and the vast majority were in common (~2.5 million). Recognizing that a single genotyping error can yield a false-positive SNP, we stratified the concordance between technologies according to minor allele frequency. As expected, the concordance is lower when the minor allele is rare; upon excluding the singleton class, both technologies identify the same SNP well more than 90% of the time.

JGIL is unique in modeling the four nucleotide frequencies rather than the frequency of reference and alternate alleles. This allows the identification of sites at which three alleles are present among the lines, as well as sites at which there exist two alleles distinct from the reference. We found 30,456 such sites in common between the two call sets, which represents >1% of all SNPs. Equally important, JGIL is able to make correct genotype calls in these unusual but not so infrequent cases. The incidence of such sites (30,456 out of 119,029,689, or 0.026%) is commensurate with the frequency of discordance between the two call sets (0.03%), suggesting that a failure to model them will substantially elevate the error rate.

Overall, the concordance across technologies between genotypes and between SNP sets suggests that JGIL is performing well. As further validation, we next compared to a pair of putative replicates on the Illumina platform, namely, lines RAL_554 and RAL_555. RAL_554 and RAL_555 appear to be replicates of one another, but this was not established until the sequencing of both was complete. Comparing the genotype calls for these two lines thus provides a measure of the error rate one should expect due to sampling and technical variability. Irrespective of quality, the overall agreement between MAP genotypes was 99.95%. To gauge the effect of our quality threshold ($Q = 20$), we classified each site according to the minimum quality of the two genotype calls being compared. We observed that for 99.75% of all sites, both genotypes were assigned $Q \geq 20$, and among these sites the calls agreed >99.99% of the time. Among the remaining 284,693 sites comprising the lower-quality class, the agreement was only 84.06%. This suggests that genotype quality scores empirically increase with the probability of the genotype being correct. To quantify this relationship, we looked at empirical quality as a function of estimated quality (again, the minimum of the two scores). The correlation between these quantifies the agreement between what is observed [empirical quality, measured as $-10\log_{10}(1 - \text{accuracy})$] and what is expected (based on JGIL posterior probability); as shown in Figure 4, that correlation is $r = 0.98$. That said, the slope of the line relating these two quantities is substantially <1 ($\hat{\beta} = 0.57$), suggesting that JGIL quality scores are overestimating the empirical quality by a predictable amount. For example, the threshold of $Q = 20$ [Pr(error) = 0.01] discussed above appears to equate to an empirical quality closer to 14 [Pr(error) = 0.04]. The strong linear relationship between the two (Fig. 4) suggests that recalibration should be possible provided a training subset of genotypes is known in advance. However, even in the absence of recalibration, it is clear that JGIL quality scores are very reliable indicators of accuracy.

The comparisons above establish the reproducibility of JGIL genotype calls across technologies and replicates. Collectively, they suggest that JGIL is highly accurate, but neither constitutes validation in the strictest sense. Thus, we turned to a third comparison based on targeted Sanger sequence data. Specifically, we obtained data for each DGRP line from five previously sequenced regions on the *X* chromosome (Arya et al. 2010). Taken together, these five regions (9,108,929–9,109,735 bp, 19,029,113–19,029,901 bp, 20,287,749–20,288,836 bp, 20,289,014–20,289,991 bp, 20,290,640–20,291,880 bp) yielded a validation set composed of nearly 5000 bp per line.

Assuming the Sanger sequence data to be correct, we sought to quantify JGIL's performance as before. There were 607,776 line/site combinations for which a Sanger call was available; the Illumina-based JGIL genotype was in agreement in 607,447 of these cases, yielding an accuracy of 99.95%. This includes JGIL calls with quality scores below 20; these were few, and their accuracy was 78.79%. It is possible for a JGIL genotype call to attain a high posterior probability in the absence of any coverage. For example, if across many lines with nonzero coverage there appears to be no variation, JGIL would predict (via its probabilistic model) that lines with no coverage likely share a genotype with the covered

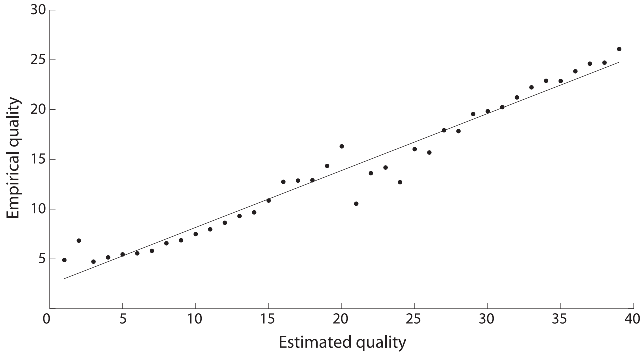

Figure 4. Empirical quality versus estimated quality. The genotype calls for replicate lines RAL_554 and RAL_555 were compared after stratification by minimum quality score (estimated quality, *x*-axis). Within each stratum, the proportion of sites for which the two calls agreed was calculated and converted to a quality (empirical quality, *y*-axis). The regression of empirical quality on estimated quality (black line) is highly significant ($R^2 = 0.9596$; $p \approx 0$). Quality scores were truncated at 40 and do not appear in the plot; the empirical quality among all sites with estimated quality $Q \geq 40$ was 47 (data not shown).

lines. The implicit assumption is that the absence of coverage is stochastic, but it is easy to imagine scenarios (e.g., a deletion) in which it is biological. We mitigate the latter case at the expense of the former by masking uncovered sites as N regardless of the probability of the MAP genotype. Between low-quality calls and uncovered sites, 6510 JGIL calls were masked in the validation set, representing close to 1% of the total. Among these, had we instead reported the MAP genotype, it would have been correct according to the Sanger data nearly 99% of the time.

Thus far we have assumed the validation set to be correct. However, because JGIL appears to be highly accurate, the Sanger error rate cannot be ignored. Because we expect this rate to be low, errors in the validation set are unlikely to distort measurements of genotype accuracy. But, because a single error in any line is enough to falsely classify a site as variable, even small error rates can strongly distort statistics regarding SNP detection. Reasoning that most errors would transform invariant sites to singleton SNPs, we restricted attention to variable sites in the validation set for which two or more lines harbored the less frequent allele. There were 90 such sites present, and JGIL identified 83 of them. This suggests a false-negative rate of just below 8%, which may in part explain why some SNPs in the initial comparison were private to one of the two technologies. False positives will also contribute, and we can quantify their incidence as well. Among the 4757 monomorphic sites in the Sanger data, JGIL identified 10 SNPs, which equates to a false-positive rate of 0.2%.

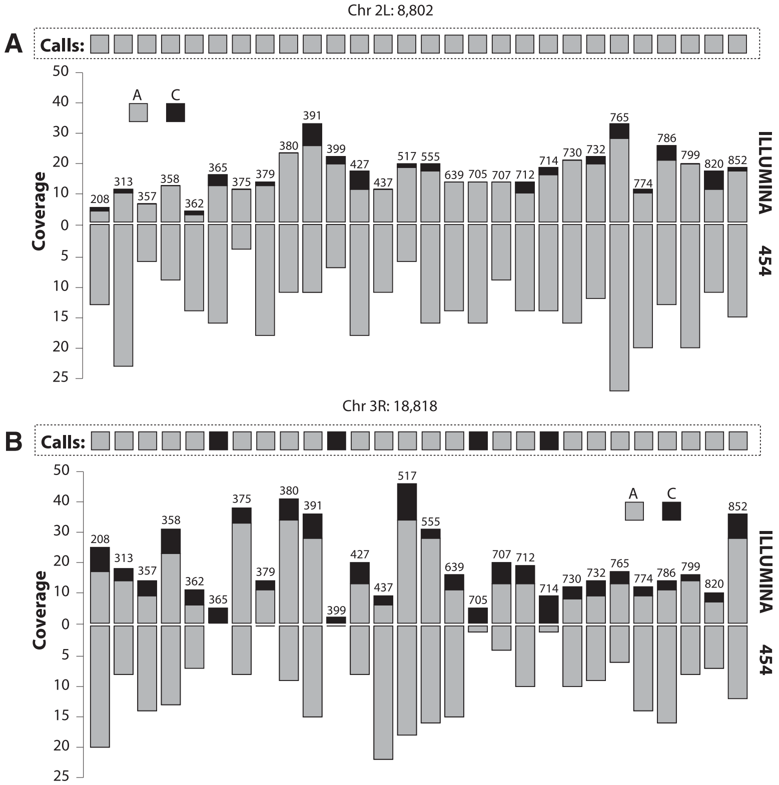

Figure 5. Two examples of JGIL applied to the DGRP. (*A*) Data for chromosome *2L*, position 8802 generated on two sequencing platforms for 28 DGRP lines. Each vertical bar summarizes the data for one line from both the Illumina GAII (oriented upward) and the Roche 454 machines (oriented downward). The height of the bar indicates coverage and is partitioned into counts of A (gray) and C (black). Each bar is labeled with the index of its corresponding line. Note that every 454 read shows an A, strongly suggesting that the Illumina C reads are erroneous. (*B*) As in panel *A* for chromosome *3R*, position 18818. Based on the Illumina data, JGIL assigns to four of the lines (365, 399, 705, 714) a homozygous C genotype, but this is not supported by the 454 data.

Each of the above comparisons is evidence that JGIL is performing well. Sometimes this performance is the result of a clear signal in the mapped reads; other times, however, JGIL is able to reason through data that are, in qualitative terms, confusing. Generally speaking, three features allow JGIL to accommodate unusual and aberrant sites. First, because JGIL estimates a site-specific allele frequency vector $\mathbf{p}$, it is able to accommodate sites at which more than two alleles are segregating within the sample. This cannot be accomplished in a biallelic model in which only ancestral and derived bases are considered. Second, because JGIL estimates a site-specific error probability vector $\boldsymbol{\varepsilon}$, it can properly genotype sites whose reads are contaminated with a low level of mapping error. An example of this is given in Figure 5A. Shown is a summary of the Illumina and 454 read data at position 8802 on chromosome *2L*, from which it is apparent that 18 of the 28 lines have Illumina coverage for a C nucleotide that is not corroborated by 454. JGIL estimates from the Illumina data that $\hat{\mathbf{p}} = (1, 0, 0, 0)$ and $\hat{\boldsymbol{\varepsilon}} = (0, 0.067, 0, 0)$, essentially reasoning out that every C is an error. For every line, the posterior probability of A is effectively 1, and JGIL calls this site as invariant.

In more extreme cases, such as that of position 18818 on chromosome *3R*, the degree of error in the data may be too much for JGIL to overcome. As Figure 5B shows, the C nucleotide is ubiquitous among the lines, and for several of the lines there is no coverage of A at all. On the other hand, whenever reads with A are present for a line, they are present in greater number than those with C. JGIL estimates from the Illumina data that $\hat{\mathbf{p}} = (0.87, 0.13, 0, 0)$ and $\hat{\boldsymbol{\varepsilon}} = (0, 0.24, 0, 0)$, indicating that there is some confusion over the presence of C (as evidenced by the dot product of $\hat{\mathbf{p}}$ and $\hat{\boldsymbol{\varepsilon}}$ being substantially positive). For the four lines that have no coverage for A (365, 399, 705, 714), JGIL makes a C call, although there is some posterior support for an A call in each case. For the remaining 24 lines, JGIL calls an A, which the 454 data suggest is probably correct. Importantly, despite the conflicting signals in the data, the method is not misled into calling rampant residual heterozygosity at this site. Thus, while the contamination is sufficient to mislead JGIL on a subset of the genotype calls, the estimated error probabilities can rescue the calls of lines for which the mapping error is lesser. This is the third feature of JGIL: Because the experimental design is modeled through $\Pr(\mathbf{G}|\mathbf{p})$, the method is less prone to calling joint genotypes that are wildly inconsistent with the expectation of 20 generations of inbreeding. JGIL will still identify rare sites at which there appears to be residual heterozygosity in many lines, but for it to do so, the signal must be strong and not more easily attributable as mapping error.

To explore these observations formally, we turned to a series of simulation studies. In what follows, we illustrate and quantify how the performance of JGIL responds to varying coverage, sequencing error, mapping error, and the number of lines in the analysis.

Simulation results

We first show how genotyping error rate decreases as the number of lines being genotyped grows. We then demonstrate the robustness of JGIL to moderate mapping error. Finally, we interrogate the utility of JGIL for population genomics by quantifying its ability to ascertain SNPs and estimate allele frequencies. Throughout this section, we use a Poisson model of coverage and uniform model of error.

We begin by illustrating the quantitative benefits of joint genotyping. Figure 6 plots genotyping error rate against sequencing error rate for nine combinations of coverage (two, five, 10) and number of lines (two, five, 10). For a given sequencing error rate and fixed coverage, it is clear that the genotyping error rate decreases as the number of lines increases. This decrease is directly attributable to the simultaneous inference of allele frequencies and error probabilities; moreover, it becomes more dramatic at higher sequencing error rates. Intuitively, JGIL borrows strength across lines to disentangle signal (i.e., genotypes) from noise (i.e., error). This feature becomes even more pronounced in the face of mapping error.

Previously, Figure 5 summarized two DGRP sites for which estimating error probabilities across lines led to robust genotype calls despite considerable mapping error. Supplemental Figure 2 presents a simulation study of the same phenomenon in which mapping error was simulated for varying numbers of lines and coverage. Supplemental Figure 2A plots the genotyping error rate observed at coverage 2. The frequency of mismapped reads was set as a percentage of the expected coverage for the true allele; for example, 50% represents an expected coverage of 1 for the erroneous base. Remarkably, even at minimal coverage, five lines appear to be sufficient to mitigate substantial mapping error (Supplemental Fig. 2B). Just as in Figure 5, JGIL is able to probabilistically identify the mismapped nucleotides and diminish their contribution to the genotype calls. For absolute levels of mapping error, e.g., comparing 50% error at coverage 2 to 20% error at coverage 5 and 10% error at coverage 10, higher coverage of the true allele leads to a lower genotyping error rate. If, however, the mismapped reads are derived from another location in the sequenced genome, then it may be more reasonable to compare across relative levels of mapping error. Here, the relationship appears to be more complicated. At least under a naive Poisson model, a high relative mapping error, say 90%, has the potential to be more problematic when the coverage is higher. Faced with the consistent signal of two nearly equally frequent alleles, JGIL will assume that there is ubiquitous residual heterozygosity at the site. This scenario occurs more often at coverage 10 than at coverage 2, leading to the seemingly counterintuitive result shown in Supplemental Figure 2.

Thus far we have focused on individual call quality by attempting to quantify genotyping error rates as a function of specified covariates. We next sought to quantify the performance of JGIL on a per-site basis rather than on a per-call basis, with the goals of SNP detection and allele frequency estimation in mind (Lynch 2009). We used simulation to characterize SNP ascertainment bias, as well as to measure any bias in allele frequency estimates, as a function of the true allele frequency, the error rate, and coverage. The first two rows of Supplemental Figure 3 summarize a study in which 100 lines were genotyped from reads of average depth two (left column), five (middle column), and 10 (right column). The top row reports how often among 10,000 replicates a SNP went undetected given a specified error rate (between 0.01 and 0.1 inclusive) and a specified minor allele frequency (between 0.01 and 0.1 inclusive, corresponding to between one and 10 lines). It is evident and not surprising that singletons are difficult to detect at low coverage. It is also clear, however, that either an increase in coverage or an increase in minor allele frequency is sufficient to allow rare SNPs to be detected. As the figure shows, this holds true even as the error rate increases. The middle row of the figure quantifies bias in allele frequency estimation under the same simulation conditions. Again, for low coverage, the bias is appreciable, but here the bias *increases* with minor allele frequency. This observation is artifactual in that, in the absence of covering reads, there is more a priori support and hence more a posteriori support for the major allele than for the minor allele. If one filters on either posterior probability or realized coverage, then the trend toward increasing bias goes away.

The bottom row of Supplemental Figure 3 contrasts the remainder of the figure in that there is assumed to be no minor allele. Thus, whereas the top row concerns false negatives, the bottom row considers false positives. Each simulation specified the expected coverage (two, five, 10), the error rate (from 0.01 to 0.10), and the number of lines to be analyzed (from five to 50); for each combination, the figure reports the proportion of 10,000 replicate simulations in which JGIL mistakenly called a SNP. It is clear that there is an appreciable false-positive rate at low coverage for moderate sequencing error. Moreover, this error rate evidently increases with the number of lines. As the number of lines grows, more chances arise for a line to be covered exclusively by erroneous alleles. The probability of such an event is higher at low coverage and higher as the error rate increases. The effect is similar to that of the mapping example in Figure 5B for lines 365, 399,

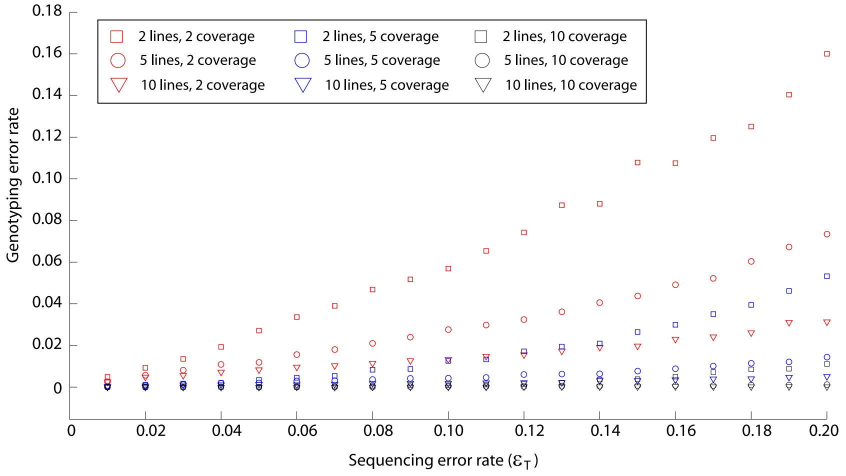

Figure 6. Genotyping error rate as a function of sequencing error rate, coverage, and sample size. At a single site, sequencing read data were simulated for either two (square), five (circle), or 10 (triangle) lines assumed to be homozygous for the nucleotide A. These were considered in combination with simulated coverage of either two reads (red), five reads (blue), or 10 reads (black). For sequencing error rates ranging from 0.01 to 0.20, the JGIL genotyping error rate was calculated across 10,000 replicate simulations.

705, and 714; if the covering reads are exclusively wrong, JGIL cannot help but be misled. On the other hand, from the figure, it appears that, for moderate coverage, sporadic sequencing error does not contribute much to the false-positive rate. But note that the same cannot be said for systematic errors (e.g., mapping or sequencing chemistry bias) such as those considered in Supplemental Figure 2.

Discussion

Improvements in sequencing technology are facilitating increasingly deep studies of genomic variation. Sometimes the unit of study is not a single individual but rather a strain within which the genetic diversity is minimal. To capture within-strain variation, it is common to pool the DNA of many individuals for sequencing, creating a mixture of genotypes that represent the strain. It was in this context that we developed JGIL to identify variation among a large panel of strains. We introduced a framework for simultaneously genotyping multiple strains by jointly analyzing their sequence reads. We showed how the nature of the data and the design of the experiment could be incorporated into a single probabilistic model. We implemented this model as JGIL for the DGRP, and we showed that it produced highly accurate genotypes in the absence of heuristics. Despite this focus, our framework addresses concerns that are general and fundamental to the sequencing of inbred lines and strains. JGIL was tailored to the DGRP only in how the probability $\Pr(\mathbf{G}|\boldsymbol{\theta})$ was structured; the number of generations can already be varied, and extensibility to other experimental designs is straightforward.

Because JGIL did not use heuristics on top of its probabilistic model, its treatment of sequence data is both sophisticated and naive. Sophisticated aspects of the method include its treatment of experimental design and its approach to modeling nucleotide frequencies and errors. These conferred several benefits in the DGRP analysis. First and foremost, JGIL was not misled into overestimating the number of sites at which each line harbors residual variation. The incidence of "heterozygotes" would have been much higher had the lines been viewed as individuals, had they been considered in isolation, or had they been considered in the absence of the experimental design. Second, JGIL was able to identify and accurately assign genotypes at sites where three alleles were present between the lines and the reference. Third, JGIL appears able to mitigate some degree of mapping error, and its parameters can be used to quantitatively flag sites where the data appears qualitatively confusing.

Other aspects of JGIL may certainly be considered naive. For example, JGIL considers each site in isolation, thereby disregarding any information encoded in neighboring sites. It is well appreciated that haplotype information generally has the potential to improve genotyping accuracy (see, e.g., Nielsen et al. 2011). On the other hand, just as the meaning of a strain genotype is complicated, so too is the meaning of a strain haplotype, and when DNA is pooled before short-read sequencing, much of the haplotype signal is lost. JGIL may also be considered naive in its ignorance of other types of genetic variation such as indels and copy number variants. Our approach takes as input the result of a reference-guided assembly, and as such its ability to ascertain genotypes is predicated on the fidelity of mapped reads. Indels complicate mapping in several ways, leading to both false negatives (e.g., proximal nucleotide variants missed because of a failure to map correctly) and false positives (e.g., artificial nucleotide variants created by reads that are incorrectly mapped). This is a limitation that JGIL shares with any method that conditions on the mapped reads; however, it may be mitigated somewhat by applying post hoc filters to flag sites where the alignment appears dubious.

We believe that the naive aspects of JGIL are more than compensated for by its sophistications. In particular, we emphasize the importance of modeling the data and the experimental design, and we advocate that when possible, $\Pr(\mathbf{G}|\boldsymbol{\theta})$ should be specified with these issues in mind. None of this is to the exclusion of complementary approaches that improve data quality and/or inference. In our application to the DGRP, for instance, we relied on the GATK package to improve the JGIL input, and we certainly envision that others will apply post hoc filters to the JGIL output. Our purpose here was to introduce a probabilistic model specifically designed to genotype a panel of strains and to show that it could be implemented to achieve highly accurate genotype calls.

Methods

Defining and encoding genotypes

We define the genotype of a line as the allelic content of its final full-sib mated pair (see Fig. 1). For example, if in one line both of these flies were AA homozygotes (as one might expect after 20 generations of inbreeding), we call and code the genotype for the line as (4,0,0,0), meaning 4 A alleles and 0 alleles of C, G, and T respectively. Alternatively, if this last generation of full-sib mating features an AC × AA cross, the genotype is then (3,1,0,0). As a simplification, we have reduced the number of possible line genotypes to 22 (see Table 1) by assuming that no more than two nucleotides will be represented in this cross for any one line. We denote the 4×22 matrix of possible genotypes in Table 1 as $\mathbf{M}$.

Sequencing and data preprocessing

Each line was sequenced on the Illumina platform to at least 12× coverage; reads ranged from 36 bp to 110 bp in length. Lines sequenced on the 454 machine had a minimum of 5× coverage with the majority above 10×, with reads ranging from 200 bp to 400 bp in length. Details for each line are given in Mackay et al. (2012). The bwa software (Li and Durbin 2009) was used to align the sequencing reads for each line to the *D. melanogaster* reference sequence 5.13 (obtained from FlyBase). The GATK package (McKenna et al. 2010) was used to recalibrate and locally realign the bam files; recalibration was seeded with a liberal list of putative variants obtained via AtlasSNP (Shen et al. 2009). Reads with a mapping quality below 10 were discarded prior to variant calling. Bases with base quality below 25 were also removed from consideration. DGRP community resources, including the lines, sequences, read alignments, and SNPs are publicly available as detailed in Mackay et al. (2012).

Modeling the inbreeding process through $\Pr(\mathbf{G}|\mathbf{p})$

Full-sib inbreeding is an iterative sampling procedure in which the genotypes at generation $n+1$ depend on those in generation n. The process follows a Markov chain whose initial state, the genotypes at generation 0, has a probability distribution specified by the population allele frequency vector $\mathbf{p}$ (Robertson 1952). For example, the probability that the G_0 cross is AA × AA is simply p_A^4, while the probability that the G_1 cross is AA × AA is $p_A^4 + p_A^3(1-p_A) + (1/4)p_A^2(1-p_A)^2$. The latter expression is more complicated because an AA × AA G_1 cross can result from three G_0 genotypic configurations: (1) AA × AA, (2) AA × AX, or (3) AX × AY, where "X" and "Y" represent any nucleotides other than A. The complexity of these probabilities coupled with our desire for an efficient implementation led us to make some benign simplifying assumptions. In particular, because the chance of any one line retaining more than two alleles at a site after 20 generations of inbreeding is negligible, we made the assumption that there exist no more than two alleles in the G_0 parentals.

Under the assumption that no more than two alleles are present in the G_0 of any one line, the number of possible genotypic configurations is 22 (enumerated in Table 1). Once the two alleles have been specified, say A and a, only six of the 22 configurations are relevant: (1) AA × AA, (2) AA × Aa, (3) AA × aa, (4) Aa × Aa, (5) Aa × aa, and (6) aa × aa. The initial probabilities of these states depend on $\mathbf{p}$, and the distribution of parental genotypes at generation $n+1$ given those at generation n can be represented by a 6 × 6 Markov transition matrix $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1/16 & 1/4 & 1/8 & 1/4 & 1/4 & 1/16 \\ 0 & 0 & 0 & 1/4 & 1/2 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This says, for example, that the probability that the G_5 cross is AA × Aa given that the G_4 cross is Aa × Aa is $\mathbf{Q}_{4,2} = 1/4$. The distribution of parental genotypes at G_{20} given the distribution at G_0 (i.e., the founding parents sampled from the population) can be found from $\mathbf{Q}^{20}$.

The Markov chain $\{X_n : n = 0, 1, \ldots, 20\}$ described by $\mathbf{Q}$ has two absorbing states, namely, state 1 (AA × AA) and state 6 (aa × aa). In other words, if the G_0 parents are homozygous for the same allele so that the Markov chain begins in an absorbing state, then the chain will remain in the same state through X_{20}. Let $\tau = \min_n\{X_n = 1 \text{ or } X_n = 6\}$ denote the first time that the chain enters an absorbing state. Then $\Pr(\tau > 20|X_0 = s)$ is the probability that residual heterozygosity remains after 20 generations of inbreeding when the G_0 parents have genotypes described by state s. It turns out that $\Pr(X_{20} = 2|\tau > 20, X_0 = s \in \{2, 3, 4, 5\}) \cong 3/10$ independent of the initial state. Similarly, the values for $X_{20} = 3, 4, 5$ are $1/20, 7/20, 3/10$, respectively. We use the full-sib inbreeding coefficient F_{20} to approximate $\Pr(\tau > 20)$ independent of the initial state. F_{20} can be calculated using the recurrence relation $F_{t+2} = 0.25 + (0.5)F_{t+1} + (0.25)F_t$, from which $F_{20} = 0.9863$ is obtained (Crow and Kimura 1970).

Using these approximations, we can calculate $\Pr(G_i = g|\mathbf{p})$ for each of the 22 genotypes shown in Table 1 through

$$\begin{aligned} \Pr(X_{20} = s|\mathbf{p}) = & \sum_{z=1}^{6} \Pr(X_{20} = s|X_0 = z, \tau > 20) \\ & \times \Pr(\tau > 20|X_0 = z) \Pr(X_0 = z|\mathbf{p}) \\ & + \sum_{z=1}^{6} \Pr(X_{20} = s|X_0 = z, \tau \le 20) \\ & \times \Pr(\tau \le 20|X_0 = z) \Pr(X_0 = z|\mathbf{p}) \end{aligned}$$

We approximate these probabilities with the $\mathbf{v}$ defined in the section entitled "Updating equations for the EM."

Modeling the sequencing process through $\Pr(\mathbf{R}|\mathbf{G}, \boldsymbol{\varepsilon})$

Recall that there are N_i reads for line i. Assuming some arbitrary order for the reads, let R_i^k denote read k for line i. Then

$$\begin{aligned} \Pr(R_i = r_i|N_i, G_i = g, \boldsymbol{\varepsilon}) &= \prod_{k=1}^{N_i} \Pr\left(R_i^k = r_i^k|G_i = g, \boldsymbol{\varepsilon}\right) \\ &= \prod_{k=1}^{N_i} \prod_{j=1}^{4} \left[\underbrace{\frac{M_{ij}}{4}(1 - \varepsilon_. + \varepsilon_j)}_{\Pr\left(\mathrm{r}_\mathrm{i}^\mathrm{k} = j, \text{ no error}\right)} + \underbrace{\left(1 - \frac{M_{ij}}{4}\right)(\varepsilon_. - \varepsilon_j)}_{\Pr\left(r_i^k = j, \text{ error}\right)} \right] \mathbf{1}_{\{r_i^k = j\}}, \end{aligned}$$

where $\varepsilon_.$ is used to denote the sum of the entries in $\boldsymbol{\varepsilon}$. Note that we have explicitly modeled the realized coverage N_i as fixed rather than random.

Data augmentation and the expected log-likelihood

We have described a probabilistic model for $\Pr(\mathbf{R}|\mathbf{p}, \boldsymbol{\varepsilon})$, where $\mathbf{R}$ is the 4 × m matrix of read counts at a site across lines and $\mathbf{p}$ and $\boldsymbol{\varepsilon}$ are the site-specific population frequencies and error probabilities, respectively. To estimate the parameters via maximum likelihood, we must find $(\hat{\mathbf{p}}, \hat{\boldsymbol{\varepsilon}}) = \operatorname{argmax}_{\mathbf{p},\boldsymbol{\varepsilon}} L(\mathbf{p}, \boldsymbol{\varepsilon}|\mathbf{R}) = \operatorname{argmax}_{\mathbf{p},\boldsymbol{\varepsilon}} \Pr(\mathbf{R}|\mathbf{p}, \boldsymbol{\varepsilon})$, which by direct maximization would be challenging. Instead, we appeal to the missing data previously described. Specifically, we take as missing data the unobserved 1 × m vector of genotypes $\mathbf{G}$ and the m unobserved indicator vectors Y^i of length 1 × N_i.

We can write

$$\Pr(\mathbf{R} = \mathbf{r}|\mathbf{p}, \boldsymbol{\varepsilon}) = \prod_{i=1}^{m} \sum_{g=1}^{22} \Pr(R_i = r_i|G_i = g, \boldsymbol{\varepsilon}) \Pr(G_i = g|\mathbf{p}).$$

Now, $\Pr(G_i = g|\mathbf{p})$ is a function of the inbreeding design, and its derivation is detailed above. The remaining term $\Pr(R_i = r_i|G_i = g, \boldsymbol{\varepsilon})$ can be expressed with the help of the read indicators as we now describe. For $k = 1, \ldots, N_i$, define

$$Y_k^i = \begin{cases} 0, & \text{if read } k \text{ of line } i \text{ has the correct base} \\ 1, & \text{if read } k \text{ of line } i \text{ has an error} \end{cases}.$$

Then upon augmenting the read data with the indicators, we have

$$\begin{aligned} \Pr(R_i = r_i, Y_i|G_i = g, \varepsilon) = & \prod_{k=1}^{N_i} \prod_{j=1}^{4} \left(\frac{M_{jg}}{4}(1 - \varepsilon_. + \varepsilon_j) \right)^{1-Y_k^i} \\ & \times \left(\left(1 - \frac{M_{jg}}{4}\right)(\varepsilon_. - \varepsilon_j) \right)^{Y_k^i} \mathbf{1}_{\{r_i^k = j\}} \end{aligned}$$

Augmenting with read indicators and genotypes, we have

$$\begin{aligned} \Pr(\mathbf{R} = \mathbf{r}, \mathbf{G}, \mathbf{Y}|\mathbf{p}, \boldsymbol{\varepsilon}) = & \prod_{i=1}^{m} \Pr(G_i = g_i|\mathbf{p}) \prod_{k=1}^{N_i} \prod_{j=1}^{4} \left(\frac{M_{jg_i}}{4}(1 - \varepsilon_. + \varepsilon_j) \right)^{1-Y_k^i} \\ & \times \left(\left(1 - \frac{M_{jg_i}}{4}\right)(\varepsilon_. - \varepsilon_j) \right)^{Y_k^i} \mathbf{1}_{\{r_i^k = j\}} \end{aligned}$$

So the augmented log-likelihood is

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{R} = \mathbf{r}, \mathbf{G}, \mathbf{Y}) = & \sum_{i=1}^{m} \log \Pr(G_i = g_i|\mathbf{p}) \\ & + \sum_{i=1}^{m} \sum_{k=1}^{N_i} \sum_{j=1}^{4} \left[(1 - Y_k^i) \log\left(\frac{M_{jg_i}}{4}(1 - \varepsilon_. + \varepsilon_j) \right) \right. \\ & \left. + Y_k^i \log\left(\left(1 - \tfrac{M_{jg_i}}{4}\right)(\varepsilon_. - \varepsilon_j) \right) \right] \mathbf{1}_{\{r_i^k = j\}} \end{aligned}$$

and for the EM we seek the parameter values for $\mathbf{p}, \boldsymbol{\varepsilon}$ that maximize its expectation with respect to the distribution $\mathbf{G}, \mathbf{Y}|\mathbf{R}, \boldsymbol{\theta}^*$. Solutions to the maximization are given below.

Updating equations for the EM

Below we describe how our estimate of $\boldsymbol{\theta}$ is updated from $\boldsymbol{\theta}^0$ to $\boldsymbol{\theta}^1$ in one iteration of the EM. Let $\mathbf{1}$ denote a vector or matrix of ones, and let $\mathbf{I}$ denote the identity matrix. We use $\mathbf{p}^0$ and $\boldsymbol{\varepsilon}^0$ to denote the previous estimates in $\boldsymbol{\theta}^0$ and treat each as a 4 × 1 column vector. The operator $\circ$ denotes the Hadamard product of two matrices.

Let F_{20} be as described above and define the 22 × 1 vector of genotypic probabilities $\mathbf{v}$ as follows. For $i = 1, \ldots, 4$, $\mathbf{v}_i = \left(\mathbf{p}_i^0\right)^2(1 - \mathrm{F}_{20}) + \mathbf{p}_i^0\mathrm{F}_{20}$. These are the homozygous probabilities based on the nucleotide frequencies in the initial population. For segregating A and C, we use $\mathbf{v}_5 = \mathbf{v}_7 = (3/5)\mathbf{p}_1^0\mathbf{p}_2^0(1 - \mathrm{F}_{20})$ and $\mathbf{v}_6 = (4/5)\mathbf{p}_1^0\mathbf{p}_2^0(1 - \mathrm{F}_{20})$; $\mathbf{v}_8$ through $\mathbf{v}_{22}$ are specified similarly for the remaining segregating pairs (see Table 1). Note that the coefficients 3/5 and 4/5 arise from the X_{20} probabilities from the Markov chain described above.

Let $\mathbf{A} = \left(\mathbf{1}_{4\times 22} - \frac{1}{4}\mathbf{M}\right)\circ\boldsymbol{\varepsilon}\mathbf{1}_{1\times 22}$ and $\mathbf{B} = \left(\frac{1}{4}\mathbf{M}\right)\circ((\mathbf{1}_{4\times 1} - (\mathbf{1}_{4\times 4} - \mathbf{I}_4)\boldsymbol{\varepsilon})\,\mathbf{1}_{1\times 22})$ be 4 × 22 matrices whose rows and columns index nucleotides (A, C, G, T) and genotypes (1–22), respectively. Let $\mathbf{J}$ be the $L \times 22$ matrix whose entries are defined as

$$\mathbf{J}_{ij} = \mathbf{v}_j \prod_{k=1}^{4} \left(\mathbf{A}_{kj} + \mathbf{B}_{kj}\right)^{\mathbf{R}_{ki}}.$$

Let $\mathbf{H}$ be the $m \times 22$ matrix obtained from $\mathbf{J}$ by normalizing each of its rows to sum to one (i.e.,

$$\mathbf{H}_{ij} = \frac{\mathbf{J}_{ij}}{\sum_{x=1}^{22}\mathbf{J}_{ix}}).$$

The m rows of $\mathbf{H}$ correspond to each of the m inbred lines and report posterior probabilities for each of the 22 genotypes (see Table 1). The 1 × 22 vector $\mathbf{1}_{1\times m}\mathbf{H}$ contains the expected number of lines of each genotype. Let $\mathbf{K}$ be the 4 × 22 matrix obtained from $\mathbf{M}$ through $\mathbf{K}_{ij} = \lceil \mathbf{M}_{ij}/3 \rceil$. Then the updated 4 × 1 vector of nucleotide frequency estimates $\mathbf{p}^*$ is given by

$$\mathbf{p}^* = \frac{\mathbf{K}\mathbf{H}^{\mathrm{T}}\mathbf{1}_{m\times 1}}{\mathbf{1}_{1\times 4}\mathbf{K}\mathbf{H}^{\mathrm{T}}\mathbf{1}_{m\times 1}}.$$

Finally, let $\mathbf{S}$ be the 4 × 22 matrix whose entries are

$$\mathbf{S}_{kj} = \begin{cases} \frac{\mathbf{A}_{kj}}{\mathbf{A}_{kj}+\mathbf{B}_{kj}}, & \text{if } \mathbf{A}_{kj} + \mathbf{B}_{kj} > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Then the updated 4 × 1 vector of error probability estimates $\boldsymbol{\varepsilon}^*$ is given by

$$\boldsymbol{\varepsilon}^* = \frac{\left(\mathbf{S}\circ(\mathbf{R}\mathbf{H})\right)\mathbf{1}_{22\times 1}}{\mathbf{1}_{1\times 4}\mathbf{R}\mathbf{1}_{m\times 1}}.$$

Software access

JGIL is available for download at http://www4.ncsu.edu/~eastone2/software.

Acknowledgments

We thank the DGRP community, and especially Trudy Mackay and Julien Ayroles, for their support and patience during the development of JGIL. Many thanks to Chris Smith for developing the JGIL executable. This work was supported in part by the National Institutes of Health grant R01GM045146.

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Arya GH, Weber AL, Wang P, Magwire MM, Negron YL, Mackay TF, Anholt RR. 2010. Natural variation, functional pleiotropy and transcriptional contexts of odorant binding protein genes in *Drosophila melanogaster*. *Genetics* **186:** 1475–1485.

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, Ferris MT, Frelinger JA, Heise M, Frieman MB, et al. 2011. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Res* **21:** 1213–1222.

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43:** 956–963.

Crow JF, Kimura M. 1970. *An introduction to population genetics theory*. Harper & Row, New York.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol* **39:** 1–38.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498.

Ding L, Wendl MC, Koboldt DC, Mardis ER. 2010. Analysis of next-generation genomic data in cancer: Accomplishments and challenges. *Hum Mol Genet* **19:** R188–R196.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186–194.

Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, et al. 2010. SNVMix: Predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26:** 730–736.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25:** 2283–2285.

Le SQ, Durbin R. 2010. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* **21:** 952–960.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27:** 2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, et al. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42:** 969–972.

Lynch M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182:** 295–301.

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482:** 173–178.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12:** 443–451.

Robertson A. 1952. The effect of inbreeding on the variation due to recessive genes. *Genetics* **37:** 189–207.

Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, et al. 2009. A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20:** 273–280.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329:** 75–78.

Received July 15, 2011; accepted in revised form February 21, 2012.